DATA-ANALYTICS

Data analytics (DA) is the process of examining data sets in order to find trends and draw conclusions about the information they contain. Increasingly data analytics is used with the aid of specialized systems and software. Data analytics technologies and techniques are widely used in commercial industries to enable organizations to make more-informed business decisions. It is also used scientists and researchers to verify or disprove scientific models, theories and hypotheses.

Data analytics initiatives can help businesses increase revenues, improve operational efficiency, optimize marketing campaigns and customer service efforts. It can also be used to respond quickly to emerging market trends and gain a competitive edge over rivals. The ultimate goal of data analytics, however, is boosting business performance. Depending on the particular application, the data that's analyzed can consist of either historical records or new information that have been processed for real-time analytics. In addition, it can come from a mix of internal systems and external data sources.

****
****

DA-1(GENERAL FUNCTIONS)(CSV FILE USED: sales.csv)

The File Contains the Introduction to Data Analytics and Basic Fumctions Used While Analysing the Data. Some General Functions are:

1. Head : The head () method in python contains only one parameter, which is n. It is an optional parameter. By setting it, we fix the number of rows we want from the DataFrame. The head () function returns n rows from the DataFrame.

2. Shape : shape() is use in pandas to give number of row/column.

3. Columns : columns is used to display the Columns in the data.

4. dtypes : dtypes is used to Print Data Type of the Columns in the Data.

5. info : info() function is used to get a concise summary of the dataframe. It comes really handy when doing exploratory analysis of the data.

6. iLoc : iloc is used to extract an element based upon the position.

****
****

DA-2(CENTRAL TENDENCY AND DISPERSION)(CSV FILE USED: sales.csv)

This File Contains the Central Tendency and Dispersion Functions such as:

1. Mean : For a data set, the arithmetic mean, also called the expected value or average, is the central value of a discrete set of numbers: specifically, the sum of the values divided by the number of values.

2. Median : The median is well-defined for any ordered (one-dimensional) data, and is independent of any distance metric. The median can thus be applied to classes which are ranked but not numerical (e.g. working out a median grade when students are graded from A to F), although the result might be halfway between classes if there is an even number of cases.

3. Mode : Mode is the value which occurs most frequently in a set of observations. For example, {6, 3, 9, 6, 6, 5, 9, 3} the Mode is 6, as it occurs most often.

4. Percentile : A percentile (or a centile) is a measure used in statistics indicating the value below which a given percentage of observations in a group of observations falls.

5. Quartile : A quartile is a type of quantile which divides the number of data points into four more or less equal parts, or quarters. The first quartile (Q1) is defined as

the middle number between the smallest number and the median of the data set.

6. Inter Quartile Range : Interquartile range (IQR), also called the midspread, middle 50%, or H-spread, is a measure of statistical dispersion, being equal to the difference between 75th and 25th percentiles, or between upper and lower quartiles, IQR = Q3 − Q1.

7. Variance : Variance (σ 2) in statistics is a measurement of the spread between numbers in a data set.

8. Standard Deviation : Standard Deviation is a measure which shows how much variation (such as spread, dispersion, spread,) from the mean exists. The standard deviation indicates a "typical" deviation from the mean.

9. Population Standard Deviation : The population standard deviation is the square root of the variance.

10. Skewness : Skewness refers to distortion or asymmetry in a symmetrical bell curve, or normal distribution, in a set of data. If the curve is shifted to the left or to the right, it is said to be skewed.

11. Box Plot : A Box plot or boxplot is a method for graphically depicting groups of numerical data through their quartiles. Box plots may also have lines extending from the boxes indicating variability outside the upper and lower quartiles, hence the terms box-and-whisker plot and box-and-whisker diagram. Outliers may be plotted as individual points.

****
****

DA-3(DISTRIBUTION)

This File Contains the Most Commonly used Distribution Functions such as:

1. Binomial Distribution : Binomial distribution with parameters n and p is the discrete probability distribution of the number of successes in a sequence of n independent experiments, each asking a yes–no question, and each with its own boolean -valued outcome: success / yes / true / one (with probability p) or failure / no / false / zero (with probability q = 1 − p).

Binomial Distribution Formula

The binomial distribution formula is for any random variable X, given by;

P(x:n,p) = nCx px (1-p)n-x

Or

P(x:n,p) = nCx px (q)n-x

Where,

n = the number of experiments,

x = 0, 1, 2, 3, 4, …,

p = Probability of Success in a single experiment,

q = Probability of Failure in a single experiment = 1 – p,

The binomial distribution formula can also be written in the form of n-Bernoulli trials, where nCx = n!/x!(n-x)!. Hence,

P(x:n,p) = n!/[x!(n-x)!].px.(q)n-x

2. Poisson Distribution:

A Poisson distribution is a probability distribution which results from the Poisson experiment. A Poisson experiment is a statistical experiment that classifies the experiment into two categories, such as success or failure. Poisson distribution is a limiting process of the binomial distribution. A Poisson random variable “x” defines the number of successes in the experiment. This distribution occurs when there are events that do not occur as the outcomes of a definite number of outcomes. Poisson distribution is used under certain conditions. They are:

Number of trials “n” tends to infinity

Probability of success “p” tends to zero

np = 1 is finite

Poisson Distribution Formula

The formula for the Poisson distribution function is given by:

$f(x) = (e^{-\lambda} \lambda x)/x!$

Where,

e is the base of the logarithm,

x is a Poisson random variable,

λ is an average rate of value.

3. Uniform Distribution : A continuous probability distribution is a Uniform distribution and is related to the events which are equally likely to occur. It is defined by two parameters, x and y, where x = minimum value and y = maximum value. It is generally denoted by u(x, y).

4. Normal Distribution : A normal distribution is an arrangement of a data set in which most values cluster in the middle of the range and the rest taper off symmetrically toward either extreme.

5. Cumulative Distribution Function : The Cumulative Distribution Function (CDF), of a real-valued random variable X, evaluated at x, is the probability function that X will take a value less than or equal to x. It is used to describe the probability distribution of random variables in a table. And with the help of these data, we can create a CDF plot in excel sheet easily.

6. Hypergeometric Distribution : The hypergeometric distribution is a discrete probability distribution that describes the probability of successes (random draws for which the object drawn has a specified feature) in draws, without replacement, from a finite population of size that contains exactly objects with that feature, wherein each draw is either a success or a failure.

7. Exponential Distribution : The exponential distribution graph is a graph of the probability density function which shows the distribution of distance or time taken

between events. The two terms used in the exponential distribution graph is lambda (λ)and x. Here, lambda represents the events per unit time and x represents the time.

****
****

DA-4(HYPOTHESIS TESTING)

WHAT IS HYPOTHESIS?

A hypothesis is a proposition that attempts to explain a set of facts in a unified way. It generally forms the basis of experiments designed to establish its plausibility. Simplicity, elegance, and consistency with previously established hypotheses or laws are also major factors in determining the acceptance of a hypothesis.

WHAT IS HYPOTHESIS TESTING?

Hypothesis testing is a form of statistical inference that uses data from a sample to draw conclusions about a population parameter or a population probability distribution.

P VALUE APPROACH :

In statistics, the p-value is the probability of obtaining results at least as extreme as the observed results of a statistical hypothesis test, assuming that the null hypothesis is correct. The p-value is used as an alternative to rejection points to provide the smallest level of significance at which the null hypothesis would be rejected. A smaller p-value means that there is stronger evidence in favor of the alternative hypothesis.

t TEST :

A t-test is a type of inferential statistic used to determine if there is a significant difference between the means of two groups, which may be related in certain features. It is mostly used when the data sets, like the data set recorded as the outcome from flipping a coin 100 times, would follow a normal distribution and may have unknown variances. A t-test is used as a hypothesis testing tool, which allows testing of an assumption applicable to a population. A t-test looks at the t-statistic, the t-distribution values, and the degrees of freedom to determine the statistical significance. To conduct a test with three or more means, one must use an analysis of variance.

****
****

DA-5(ANOVA)(EXCEL FILE USED: ANOVA_ONEWAY.xlsx)

What is Analysis of Variance (ANOVA)?

Analysis of variance (ANOVA) is an analysis tool used in statistics that splits an observed aggregate variability found inside a data set into two parts: systematic factors and random factors. The systematic factors have a statistical influence on the given data set, while the random factors do not. Analysts use the ANOVA test to determine the influence that independent variables have on the dependent variable in a regression study.

The Formula for ANOVA is:

F= MSE/MST

where:

F=ANOVA coefficient,

MST=Mean sum of squares due to treatment,

MSE=Mean sum of squares due to error

****
****

DA-6(RANDOMISED BLOCK DESIGN FOR AIR TRAFFIC CONTROLLER STRESS TEST)(EXCEL FILE USED: RBD.xlsx)

A randomized block design is a type of experiment where participants who share certain characteristics are grouped together to form blocks, and then the treatment (or intervention) gets randomly assigned within each block.

Limitations of the randomized block design:

Here are some of the limitations of the randomized block design and how to deal with them:

1. It cannot deal with cases with more than 1 nuisance variable

When you want to block on 2 nuisance variables, for example, use a Latin square design.

2. Difficulty in choosing the number of blocks

Decide on the number of blocks according to expert knowledge or previous literature.

By choosing fewer blocks than you need, you may have a hard time maintaining homogeneity within each block.

Also, if you choose a larger number of blocks than you should, you may end up with fewer than enough participants in a given block to be properly randomized to receive or not the treatment.

3. Difficulty in detecting/measuring the nuisance variable

****
****

DA-7(ANOVA TWO WAY)(EXCEL FILE USED: ANOVA_TWOWAY.xlsx)

What Is a 2-Way ANOVA?

ANOVA stands for analysis of variance and tests for differences in the effects of independent variables on a dependent variable. A two-way ANOVA test is a statistical test used to determine the effect of two nominal predictor variables on a continuous outcome variable. A two-way ANOVA tests the effect of two independent variables on a dependent variable. A two-way ANOVA test analyzes the effect of the independent variables on the expected outcome along with their relationship to the outcome itself. Random factors would be considered to have no statistical influence on a data set, while systematic factors would be considered to have statistical significance. By using ANOVA, a researcher is able to determine whether the variability of the outcomes is due to chance or to the factors in the analysis. ANOVA has many applications in finance, economics, science, medicine, and social science.

ANOVA vs. 2-Way ANOVA

There are two main types of analysis of variance: one-way (or unidirectional) and two-way (bidirectional). One-way or two-way refers to the number of independent variables in your analysis of variance test. A one-way ANOVA evaluates the impact of a sole factor on a sole response variable. It determines whether the observed differences between the means of independent (unrelated) groups are explainable by chance alone, or whether there are any statistically significant differences between groups.

A two-way ANOVA is an extension of the one-way ANOVA. With a one-way, you have one independent variable affecting a dependent variable. With a two-way ANOVA, there are two independents. For example, a two-way ANOVA allows a company to compare worker productivity based on two independent variables, such as department and gender. It is utilized to observe the interaction between the two factors. It tests the effect of two factors at the same time.

****
****

DA-8(LINEAR REGRESSION-1)(EXCEL FILE USED: LinearRegression-1.xlsx)

Linear Regression

Linear regression strives to show the relationship between two variables by applying a linear equation to observed data. One variable is supposed to be an independent variable, and the other is to be a dependent variable. For example, the weight of the person is linearly related to his height. Hence this shows a linear relationship between the height and weight of the person. As the height is increased, the weight of the person also gets increased.

It is not necessary that here one variable is dependent on others, or one causes the other, but there is some critical relationship between the two variables. In such cases, we use a scatter plot to imply the strength of the relationship between the variables. If there is no relation or linking between the variables, the scatter plot does not indicate any increasing or decreasing pattern. For such cases, the linear regression design is not beneficial to the given data.

Linear Regression Equation

The measure of the extent of the relationship between two variables is shown by the correlation coefficient. The range of this coefficient lies between -1 to +1. This coefficient shows the strength of the association of the observed data for two variables.

A linear regression line equation is written in the form of:

$Y = a + bX$

where X is the independent variable and plotted along the x-axis

Y is the dependent variable and plotted along the y-axis

The slope of the line is b, and a is the intercept (the value of y when $x = 0$).

Linear Regression Formula

Linear regression shows the linear relationship between two variables. The equation of linear regression is similar to the slope formula what we have learned before in earlier classes such as linear equations in two variables. It is given by;

$Y= a + bX$

Now, here we need to find the value of the slope of the line, b, plotted in scatter plot and the intercept, a.

Simple Linear Regression

The very most straightforward case of a single scalar predictor variable x and a single scalar response variable y is known as simple linear regression. The equation for this regression is represented by;

$y=a+bx$

The expansion to multiple and vector-valued predictor variables is known as multiple linear regression, also known as multivariable linear regression. The equation for this regression is represented by;

$Y = a+bX$

Almost all real-world regression patterns include multiple predictors, and basic explanations of linear regression are often explained in terms of the multiple regression form. Note that, though, in these cases, the dependent variable y is yet a scalar.

Properties of Linear Regression

For the regression line where the regression parameters b0 and b1 are defined, the properties are given as:

1. The line reduces the sum of squared differences between observed values and predicted values.

2. The regression line passes through the mean of X and Y variable values

3. The regression constant (b0) is equal to y-intercept the linear regression

4. The regression coefficient (b1) is the slope of the regression line which is equal to the average change in the dependent variable (Y) for a unit change in the independent variable (X).

****
****

DA-9(RESIDUAL ANALYSIS)(EXCEL FILE USED: ResidualAnalysis.xlsx)

Residual analysis

1. The analysis of residuals plays an important role in validating the regression model.

2. If the error term in the regression model satisfies the four assumptions noted earlier, then the model is considered valid. Since the statistical tests for significance are also based on these assumptions, the conclusions resulting from these significance tests are called into question if the assumptions regarding ε are not satisfied.

3. The ith residual is the difference between the observed value of the dependent variable, yi, and the value predicted by the estimated regression equation, $\hat{y}i$.

4. These residuals, computed from the available data, are treated as estimates of the model error, ε.

5. As such, they are used by statisticians to validate the assumptions concerning ε. Good judgment and experience play key roles in residual analysis.

6. Graphical plots and statistical tests concerning the residuals are examined carefully by statisticians, and judgments are made based on these examinations.

7. The most common residual plot shows $\hat{y}$ on the horizontal axis and the residuals on the vertical axis.

8. If the assumptions regarding the error term, ε, are satisfied, the residual plot will consist of a horizontal band of points.

9. If the residual analysis does not indicate that the model assumptions are satisfied, it often suggests ways in which the model can be modified to obtain better results.

****
****

DA-10(DATA TRUCKING)(EXCEL FILE USED: Trucking.xlsx)

Linear regression is a simple but powerful tool to analyze relationship between a set of independent and dependent variables. But, often people tend to ignore the assumptions of OLS before interpreting the results of it. Therefore, it is an essential step to analyze various statistics revealed by OLS.

In statistics, model selection is an art. a lot of factors are taken into consideration in case making this art meaningful. Let look at each of the statistic one by one and see how can that affect the reliability of the results .

1. R-squared: It signifies the "percentage variation in dependent that is explained by independent variables". Here, 73.2% variation in y is explained by X1, X2, X3, X4 and X5. This statistic has a drawback, it increases with the number of predictors(dependent variables) increase. Therefore, it becomes inconclusive in case when it is to be decided whether additional variable is adding to the predictability power of the regression.

2. Adj. R-squared: This is the modified version of R-squared which is adjusted for the number of variables in the regression. It increases only when an additional variable adds to the explanatory power to the regression.

3. Prob(F-Statistic): This tells the overall significance of the regression. This is to assess the significance level of all the variables together unlike the t-statistic that measures it for individual variables. The null hypothesis under this is "all the regression coefficients are equal to zero". Prob(F-statistics) depicts the probability of null hypothesis being true. As per the above results, probability is close to zero. This implies that overall the regressions is meaningful.

4. AIC/BIC: It stands for Akaike's Information Criteria and is used for model selection. It penalizes the errors mode in case a new variable is added to the regression equation. It is calculated as number of parameters minus the likelihood of the overall model. A lower AIC implies a better model. Whereas, BIC stands for Bayesian information criteria and is a variant of AIC where penalties are made more severe.

5. Prob(Omnibus): One of the assumptions of OLS is that the errors are normally distributed. Omnibus test is performed in order to check this. Here, the null

hypothesis is that the errors are normally distributed. Prob(Omnibus) is supposed to be close to the 1 in order for it to satisfy the OLS assumption. In this case Prob(Omnibus) is 0.062, which implies that the OLS assumption is not satisfied. Due to this, the coefficients estimated out of it are not Best Linear Unbiased Estimators(BLUE).

6. Durbin-watson: Another assumption of OLS is of homoscedasticity. This implies that the variance of errors is constant. A value between 1 to 2 is preferred. Here, it is ~1.8 implying that the regression results are reliable from the interpretation side of this metric.

7. Prob(Jarque-Bera): It i in line with the Omnibus test. It is also performed for the distribution analysis of the regression errors. It is supposed to agree with the results of Omnibus test. A large value of JB test indicates that the errors are not normally distributed.

****
****

DA-10(MAXIMUM LIKELIHOOD ESTIMATION)(EXCEL FILE USED: Mle.xlsx)

1. Maximum likelihood estimation (MLE) is a technique used for estimating the parameters of a given distribution, using some observed data.

2. For example, if a population is known to follow a normal distribution but the mean and variance are unknown, MLE can be used to estimate them using a limited sample of the population, by finding particular values of the mean and variance so that the observation is the most likely result to have occurred.

3. MLE is useful in a variety of contexts, ranging from econometrics to MRIs to satellite imaging. It is also related to Bayesian statistics.

****
****

DA-11(LOGISTIC REGRESSION)(EXCEL FILE USED: Mle.xlsx)

DA-10(LOGISTIC REGRESSION - FULL)(EXCEL FILE USED: Mle.xlsx)

Logistic regression is a classification algorithm. It is used to predict a binary outcome based on a set of independent variables.

Ok, so what does this mean? A binary outcome is one where there are only two possible scenarios—either the event happens (1) or it does not happen (0). Independent variables are those variables or factors which may influence the outcome (or dependent variable).

So: Logistic regression is the correct type of analysis to use when you're working with binary data. You know you're dealing with binary data when the output or dependent variable is dichotomous or categorical in nature; in other words, if it fits into one of two categories (such as "yes" or "no", "pass" or "fail", and so on).

However, the independent variables can fall into any of the following categories:

Continuous—such as temperature in degrees Celsius or weight in grams. In technical terms, continuous data is categorized as either interval data, where the intervals between each value are equally split, or ratio data, where the intervals are equally split and there is a true or meaningful "zero". For example, temperature in degrees Celsius would be classified as interval data; the difference between 10 and 11 degrees C is equal to the difference between 30 and 31 degrees C, but there is no true zero—a temperature of zero degrees does not mean there is "no temperature". On the other hand, weight in grams would be classified as ratio data; it has the equal intervals and a true zero. In other words, if something weighs zero grams, it truly weighs nothing.

Discrete, ordinal—that is, data which can be placed into some kind of order on a scale. For example, if you are asked to state how happy you are on a scale of 1-5, the points on the scale represent ordinal data. A score of 1 indicates a lower degree of happiness than a score of 5, but there is no way of determining the numerical value between each of the points on the scale. Ordinal data is the kind of data you might get from a customer satisfaction survey.

Discrete, nominal—that is, data which fits into named groups which do not represent any kind of order or scale. For example, eye color may fit into the categories "blue", "brown", or "green", but there is no hierarchy to these categories.

So, in order to determine if logistic regression is the correct type of analysis to use, ask yourself the following:

Is the dependent variable dichotomous? In other words, does it fit into one of two set categories? Remember: The dependent variable is the outcome; the thing that you're measuring or predicting.

Are the independent variables either interval, ratio, or ordinal? See the examples above for a reminder of what these terms mean. Remember: The independent variables are those which may impact, or be used to predict, the outcome.

In addition to the two criteria mentioned above, there are some further requirements that must be met in order to correctly use logistic regression. These requirements are known as "assumptions"; in other words, when conducting logistic regression, you're assuming that these criteria have been met. Let's take a look at those now.

Logistic regression assumptions

1. The dependent variable is binary or dichotomous—i.e. It fits into one of two clear-cut categories. This applies to binary logistic regression, which is the type of logistic regression we've discussed so far. We'll explore some other types of logistic regression in section five.

2. There should be no, or very little, multicollinearity between the predictor variables—in other words, the predictor variables (or the independent variables) should be independent of each other. This means that there should not be a high correlation between the independent variables. In statistics, certain tests can be used to calculate the correlation between the predictor variables; if you're interested in learning more about those, just search "Spearman's rank correlation coefficient" or "the Pearson correlation coefficient."

3. The independent variables should be linearly related to the log odds. If you're not familiar with log odds, we've included a brief explanation below.

4. Logistic regression requires fairly large sample sizes—the larger the sample size, the more reliable (and powerful) you can expect the results of your analysis to be.

****
****

DA-11(REGRESSION ANALYSIS)(EXCEL FILE USED: RegressionAnalysis.xlsx)

Regression analysis is a set of statistical methods used for the estimation of relationships between a dependent variable and one or more independent variables. It can be utilized to assess the strength of the relationship between variables and for modeling the future relationship between them.

Regression analysis includes several variations, such as linear, multiple linear, and nonlinear. The most common models are simple linear and multiple linear. Nonlinear regression analysis is commonly used for more complicated data sets in which the dependent and independent variables show a nonlinear relationship.

Regression analysis offers numerous applications in various disciplines, including finance.

Regression Analysis – Linear model assumptions

Linear regression analysis is based on six fundamental assumptions:

1. The dependent and independent variables show a linear relationship between the slope and the intercept.

2. The independent variable is not random.

3. The value of the residual (error) is zero.

4. The value of the residual (error) is constant across all observations.

5. The value of the residual (error) is not correlated across all observations.

6. The residual (error) values follow the normal distribution.

Regression analysis in finance

Regression analysis has several applications in finance. For example, the statistical method is fundamental to the Capital Asset Pricing Model (CAPM). Essentially, the CAPM equation is a model that determines the relationship between the expected return of an asset and the market risk premium.

The analysis is also used to forecast the returns of securities, based on different factors, or to forecast the performance of a business. Learn more forecasting methods in CFI's Budgeting and Forecasting Course!

****
****

DA-12(K MEANS CLUSTERING)(EXCEL FILE USED: KMeansClustering.xlsx)

K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabeled dataset into different clusters. Here K defines the number of pre-defined clusters that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on.

It is an iterative algorithm that divides the unlabeled dataset into k different clusters in such a way that each dataset belongs only one group that has similar properties.

It allows us to cluster the data into different groups and a convenient way to discover the categories of groups in the unlabeled dataset on its own without the need for any training.

It is a centroid-based algorithm, where each cluster is associated with a centroid. The main aim of this algorithm is to minimize the sum of distances between the data point and their corresponding clusters.

The algorithm takes the unlabeled dataset as input, divides the dataset into k-number of clusters, and repeats the process until it does not find the best clusters. The value of k should be predetermined in this algorithm.

The k-means clustering algorithm mainly performs two tasks:

1. Determines the best value for K center points or centroids by an iterative process.

2. Assigns each data point to its closest k-center. Those data points which are near to the particular k-center, create a cluster.

Hence each cluster has datapoints with some commonalities, and it is away from other cluster

****
****

DA-12(CLASSIFICATION & REGRESSION TREE)(EXCEL FILE USED: CART.xlsx)

Classification And Regression Trees for Machine Learning

1. Decision Trees are an important type of algorithm for predictive modeling machine learning.

2. The classical decision tree algorithms have been around for decades and modern variations like random forest are among the most powerful techniques available.

3. Classification and Regression Trees or CART for short is a term introduced by Leo Breiman to refer to Decision Tree algorithms that can be used for classification or regression predictive modeling problems.

4. Classically, this algorithm is referred to as "decision trees", but on some platforms like R they are referred to by the more modern term CART.

5. The CART algorithm provides a foundation for important algorithms like bagged decision trees, random forest and boosted decision trees.

6. CART Model Representation

The representation for the CART model is a binary tree.

This is your binary tree from algorithms and data structures, nothing too fancy. Each root node represents a single input variable (x) and a split point on that variable (assuming the variable is numeric).

The leaf nodes of the tree contain an output variable (y) which is used to make a prediction.

Given a dataset with two inputs (x) of height in centimeters and weight in kilograms the output of sex as male or female, below is a crude example of a binary decision tree (completely fictitious for demonstration purposes only).

SUMMARY

1. The classical name Decision Tree and the more Modern name CART for the algorithm.

2. The representation used for CART is a binary tree.

3. Predictions are made with CART by traversing the binary tree given a new input record.

4. The tree is learned using a greedy algorithm on the training data to pick splits in the tree.

5. Stopping criteria define how much tree learns and pruning can be used to improve a learned tree.

****
****